Robust Wavelet Models for Event Detection in Time Series Databases

Alexander W Blocker Pavlos Protopapas

3 November, 2009

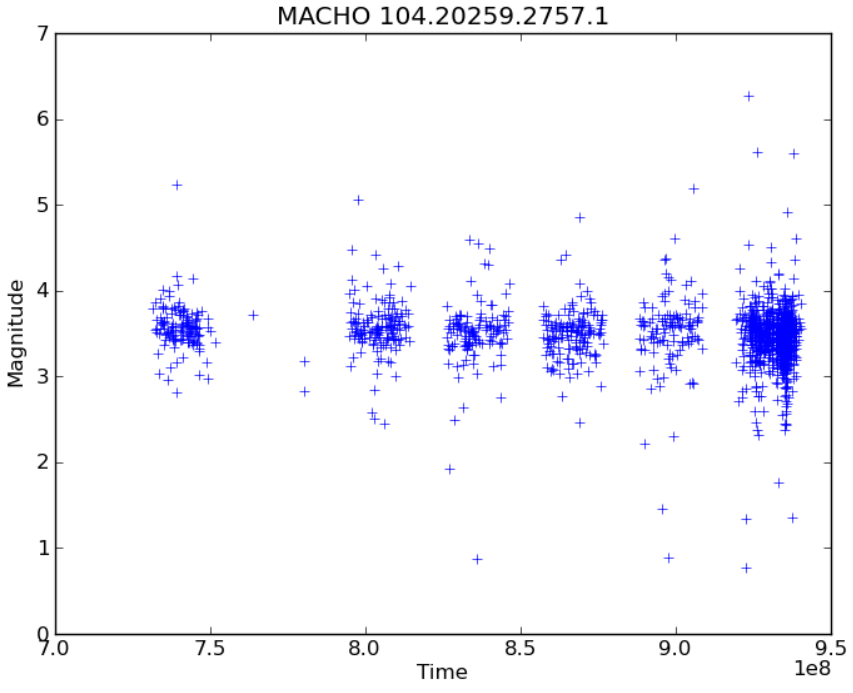
Outline

Event Detection

- We have large databases of time series (in the range of 10^5 to 10^7).
- Our goal is to identify and characterize time series containing events.
- How do we define an event?
  - We are not interested in isolated outliers. This differentiates our problem from traditional "anomaly detection" approaches.
  - We are looking for groups of observations that differ significantly from those nearby.
  - We are also attempting to distinguish quasi-periodic time series from isolated events.
- In general, taking time series to be independent observations (different sources)
  - Can sometimes combine information across series for better estimation (e.g. TAOS)

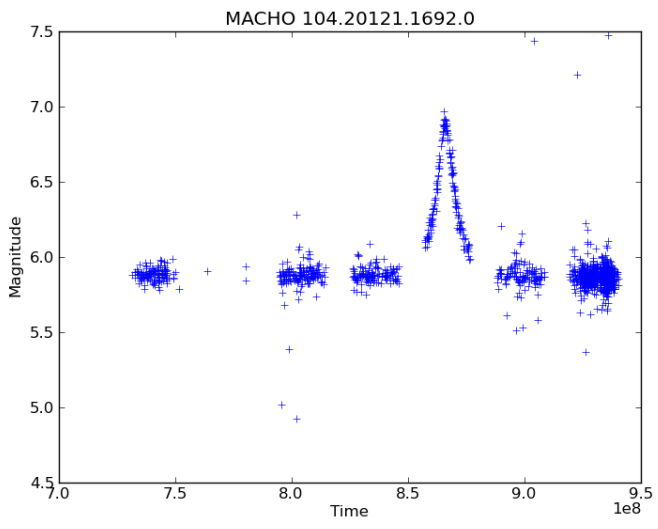
Exemplar time series from the MACHO project:

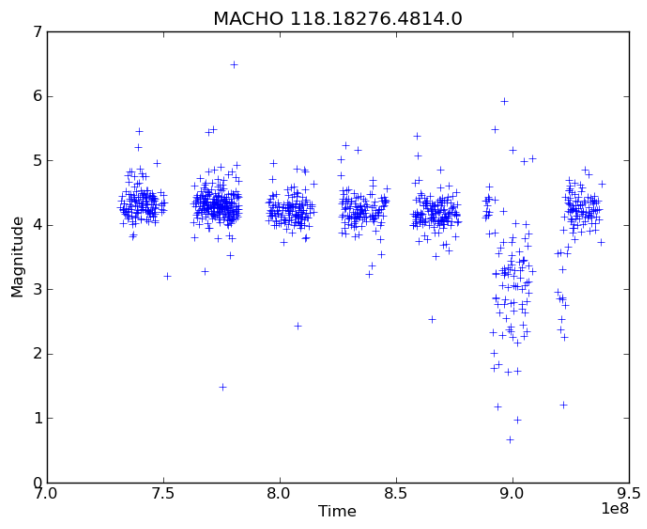
A null time series:



Exemplar time series from the MACHO project:

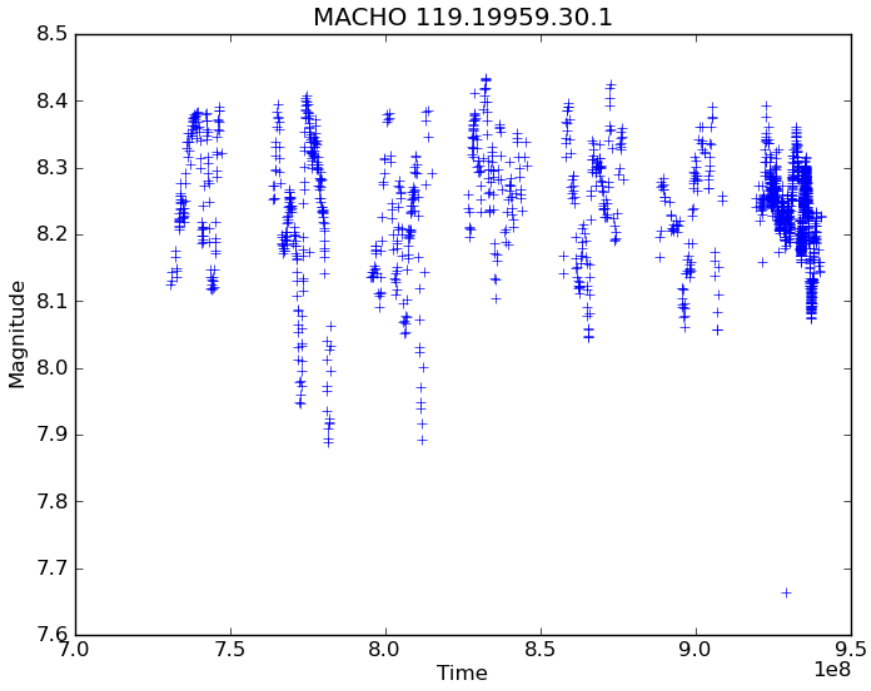
Two events:





Exemplar time series from the MACHO project:

A quasi-periodic time series:



Features of our data

- Fat-tailed measurement errors
  - Common in astronomical data, especially from ground-based telescopes
  - Requires more sophisticated modelling of data than Gaussian approaches.
- Quasi-periodic sources
  - Changes problem from binary classification to k-class
  - Requires more complex test statistics
- Non-linear, low-frequency trends complicate the definition of a baseline against which to compare our events.

Previous approaches to event detection

- Scan statistics are a common approach (Liang et al, 2004; Preston & Protopapas, 2009).
  - Typically performed on ranked data $\rightarrow$ loss of intensity information.
  - Do not deal with non-linear trends
- Equivalent width methods - common in astrophysics
  - Similar to problem of looking for lines in spectra
  - Set baseline for large window, then search for significant deviations in smaller windows
  - Loses power to multiple testing issues within each time series
  - Relies on Normal assumptions for testing

Concept

Our proposed method can be thought of as asking each time series 3 questions:

1. Is there variability?
2. If so, is it at the time scale we are interested in?
3. If it's at the right scale, is it quasi-periodic or isolated?

Notes

Some initial notes on our method:

- We do not account for irregularly spaced observations with our methods.
- We will comment on extensions to our method to handle this issue in our conclusions.

CUSUM definition

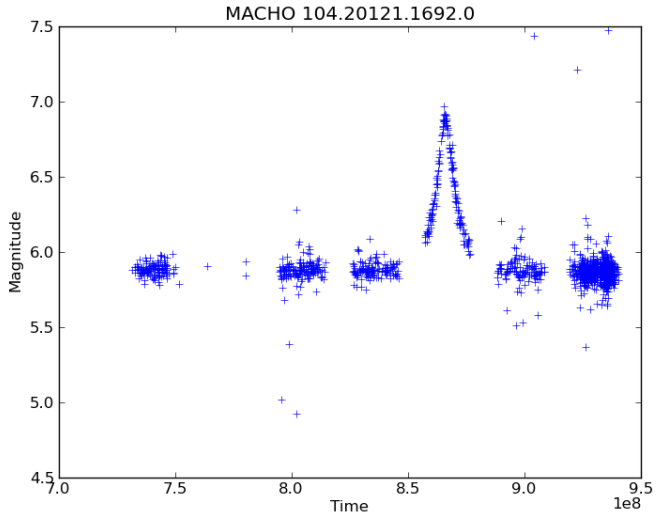
- Our first test is a simple CUSUM test.
- These have a long history in change-point detection in industrial statistics and econometrics (e.g. Ploberger, 1992, Page 1954).
- The test statistic is the range of the cumulative sum of deviations from the mean (or from a fitted linear trend):

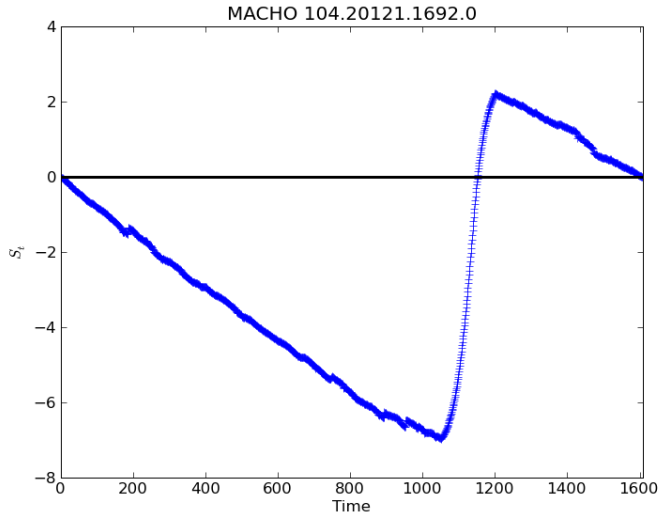
$$S_t = \frac{1}{\hat{\sigma}\sqrt{T}} \sum_{j=0}^{t} (Y_j - \hat{Y}_j)$$
$$R = \max_t(S_t) - \min_t(S_t)$$

- For T large, and assuming Gaussian residuals, the distribution of R can be approximated by the distribution of the range of a Brownian bridge.

CUSUM example

Here is one of our previous exemplar time series with its corresponding CUSUM:





CUSUM discussion

- We use the range of our CUSUM series as our statistic (as opposed to its maximum or minimum) because we are not making an assumption as to the direction of any event in our time series.
- Our CUSUM test acts as an intial screening device.
  - It would be statistically valid simply to run our final model on every time series in our database, but this would not be computationally feasible.
  - The CUSUM test is an effective way to reduce the number of time series that our computationally intensive model needs to be run on, as it has high power and, with fat-tailed data (and an incorrect Gaussian assumption for the CUSUM), is very conservative.

Model specification

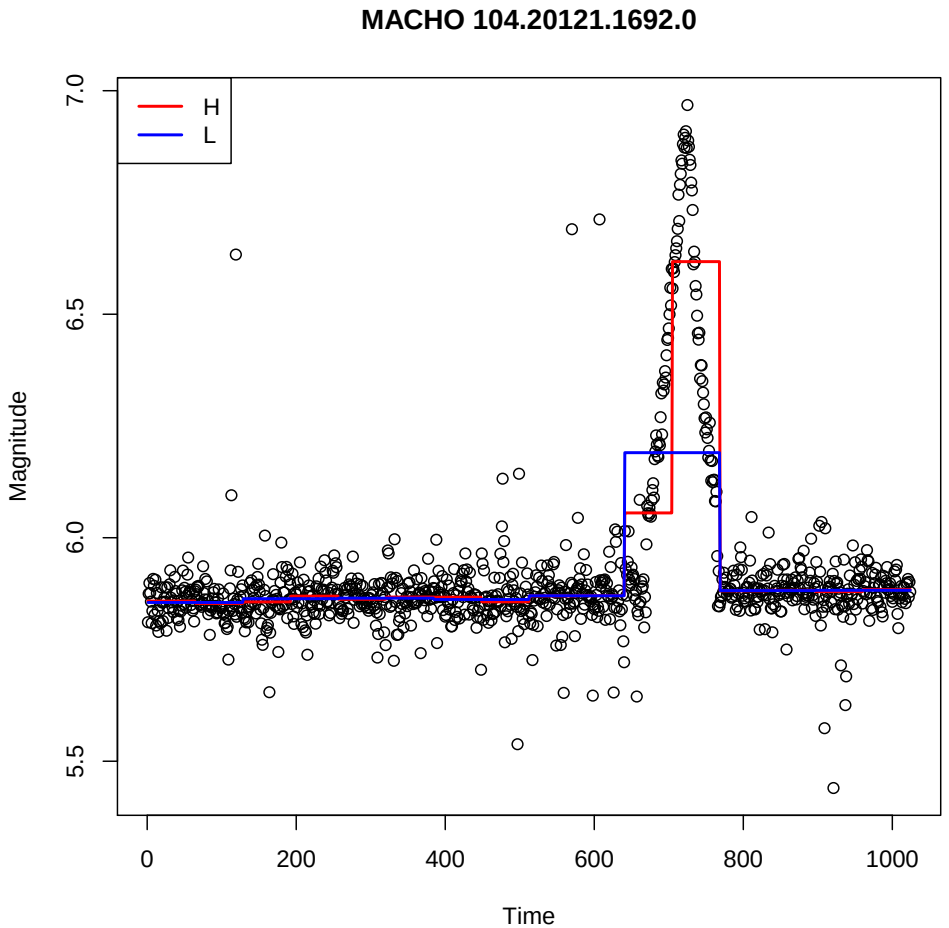
- We assume a linear model for our observations:
$$Y = X\beta + u$$
- We assume that our residuals u_t are distributed as iid $t_\nu(0, \sigma^2)$ random variables.
  - This accounts for the extreme outliers observed in our time series.
  - We typically assume a small value for ν; our default is 1, which is the most conservative for our purposes. Values between 1 and 5 typically appear reasonable for our data.
- We take X to be an incomplete wavelet basis (containing only a subset of the basis vectors).
  - Wavelets bases are ideal for our purpose because they provide discrimination in both location and scale. This is ideal for characterizing events.
  - For our purposes, we take this to be a Haar basis, but other bases may be more efficient for identifying events.

Model specification, continued

- To use this model, we first interpolate our data to a power-of-two length; call this n.
- We estimate this model for two choices of X, X_H and X_L. Each is an incomplete wavelet basis, and they are chosen to differ by one level of coefficients (e.g. X_H would be $(n \times 128)$ and X_L would be $(n \times 64)$).
- The dimensions of X_H and X_L are chosen based on prior knowledge about the events of interest
  - Such knowledge is necessary in general for event detection in the presence of trends.
  - We choose the dimension of X_L such the scale of its finest wavelet is wider than the events of interest. This choice is constrained by the requirement that the scale of the finest wavelet in X_H is narrower than the events of interest.

Example of model fit

The idea is that, if there is an event at the scale of interest, there will be a large discrepancy between the residuals using X_H and X_L:



Estimation & testing

- We fit the model specified above with each incomplete basis (X_H and X_L) using a parameter-expanded EM algorithm.
  - For details of our parameter expansion scheme, see Gelman et al. 2003
  - This algorithm is quite efficient and typically converges within 20 iterations.
- We calculate a log-likeihood ratio from our model fits:

$$\text{LLR} = 2(\ell_H - \ell_L)$$

- While this is distributed approximately χ^2 for high n and sufficiently large ν ($\nu \geq 3$), we must simulate to obtain better estimates of the relevant quantiles.

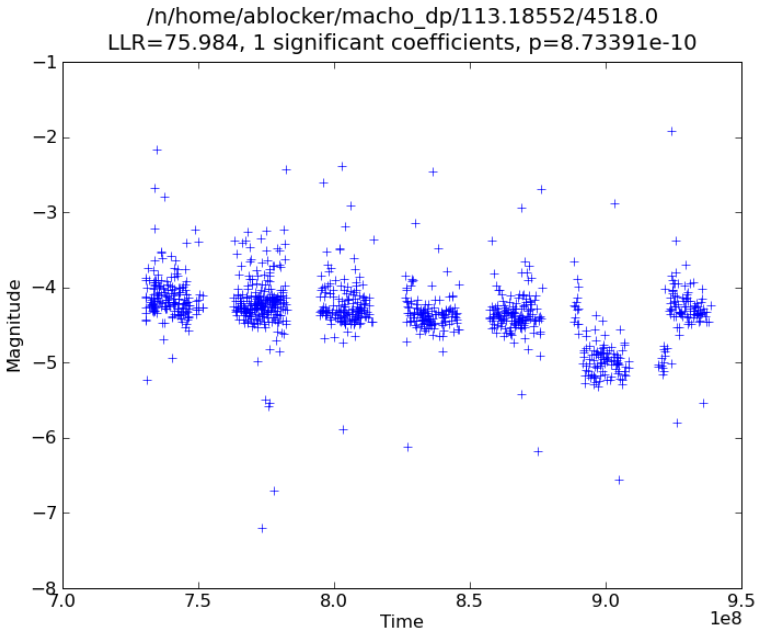
Distinguishing quasi-periodic and isolated events

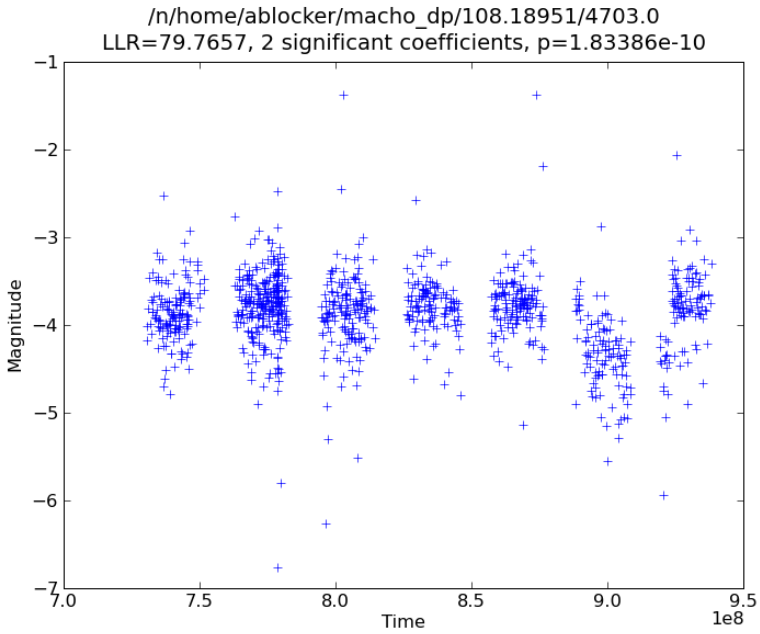
- Using the robust wavelet model described previous, the identification of quasi-periodic events is relatively straightforward.
- From each estimation with this model, we obtain a MLE for β; denote this $\hat{\beta}$.
- We are particularly interested in the finest level of coefficients from our high-resolution model (using X_H). We denote the vector of MLEs for these coefficients $\hat{\beta}_D$.
- If the time series has only one or two isolated events, we would expect relatively few of these coefficients to be significantly different from zero.
- Therefore, after testing for an event with the LLR given previously, we categorize detected time series by the number of coefficients with t-statistics exceeding some threshold (based on a Gaussian approximation to the likelihood).

Data and tuning parameter descriptions

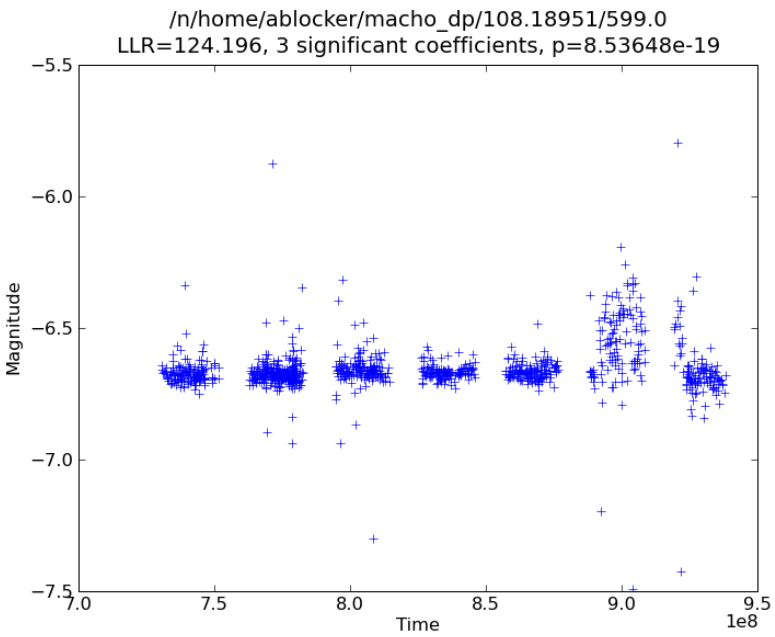
- Our subset of the MACHO database contains $515,136$ time series with lengths ranging from approximately $1,000$ to $2,000$ points.
- In stage 1, we cut all time series with $p > 10^{-4}$, leaving $79,961$ time series for stages 2 and 3.
- This leaves $29,232$ significant time series with $p \leq 10^{-7}$, including both isolated and quasiperiodic events. Of these, $4,307$ have 4 or less significant t-statistics, indicating probable isolated events.

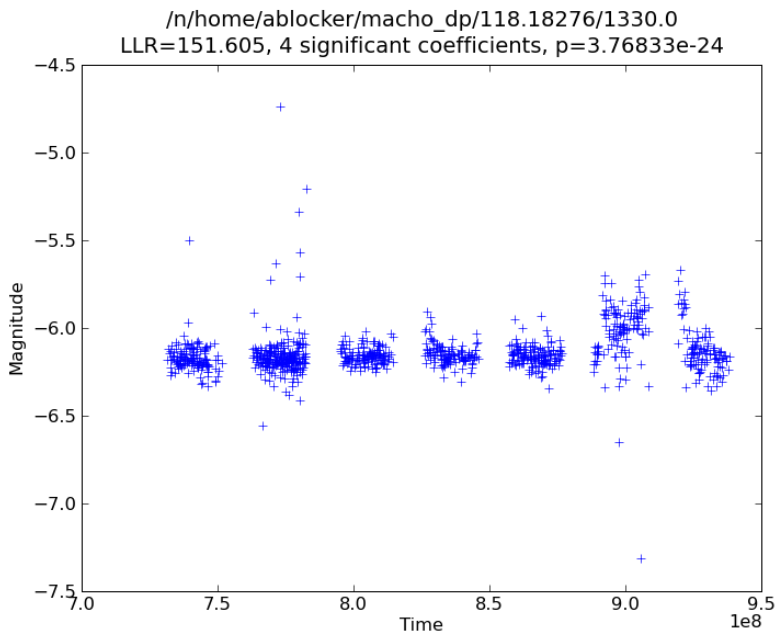
Examples of time series randomly drawn from selected categories



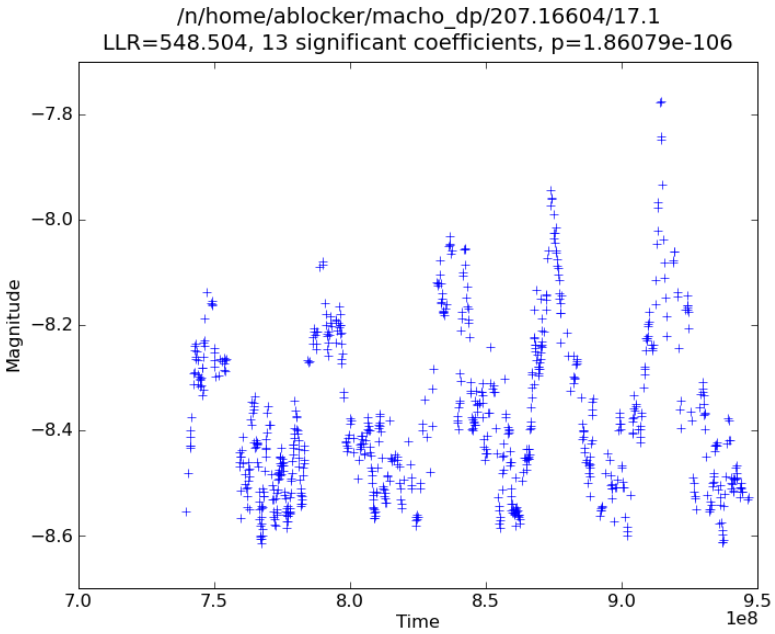


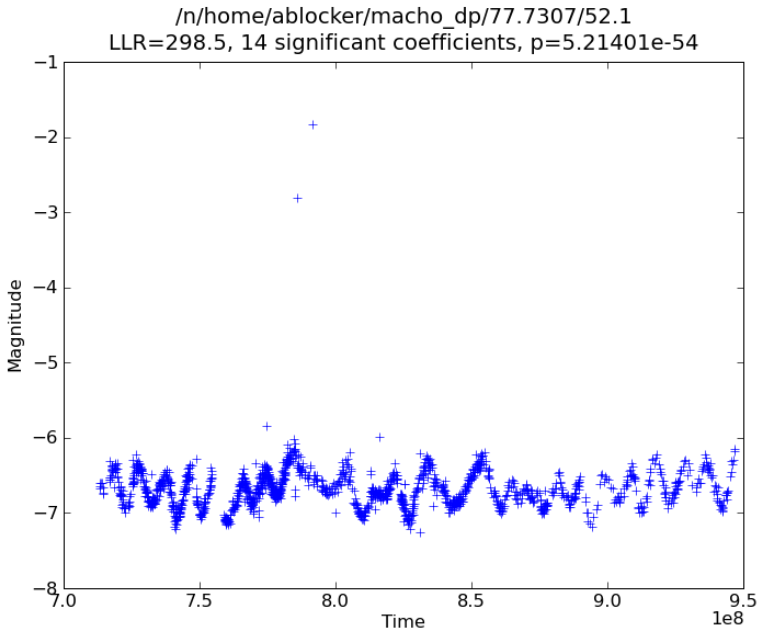
Examples of time series randomly drawn from selected categories



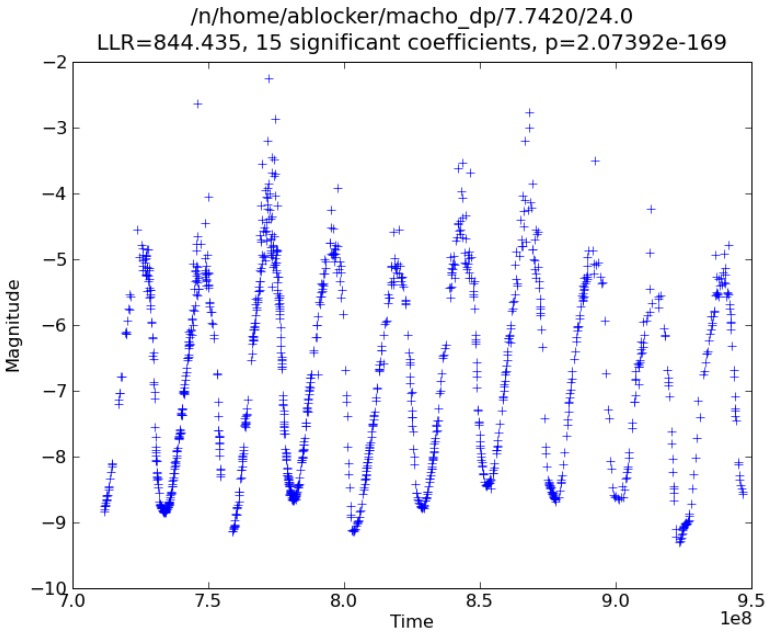


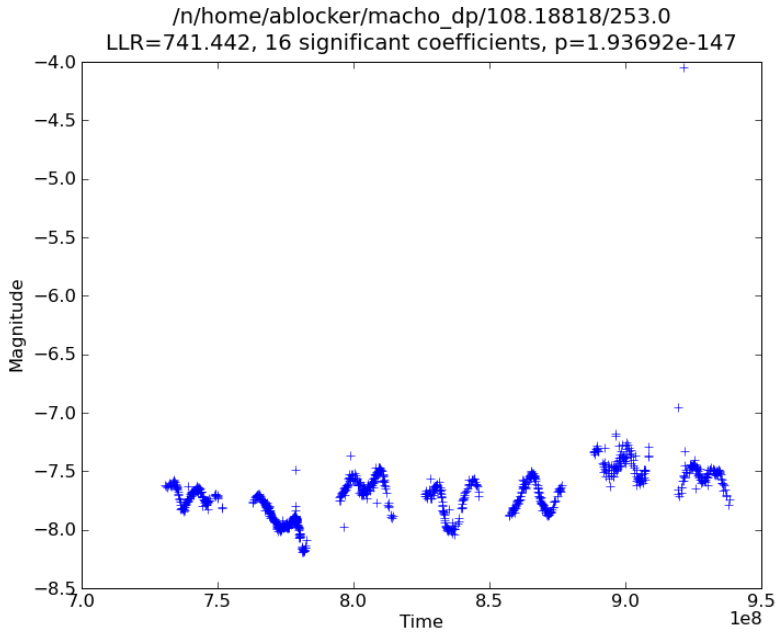
Examples of time series randomly drawn from selected categories





Examples of time series randomly drawn from selected categories





Performance on known events

- All 5 known blue stars (quasi-periodic) were found.
- 43 of 63 known microlensing events were found.
  - Why so few? Visual inspection revealed two characteristics of the missed events. First, they tended to be narrower than our high-resolution basis. Second, they were often located in time series with high levels of noise.
  - Fitting with a finer basis should resolve these misses.
- 3 of 57 known variable stars were declared significant in stage 2; only 10 passed stage 1.
  - Upon visual inspection of missed variable stars, this was unsurprising; undetected variables showed no clear signs of structured variability.

Further questions

- We currently set our p-value thresholds to control FWER; would it be more fruitful to control FDR instead?
- The Haar basis is a simple, easy to understand choice. Would other wavelets (such as least-asymmetric Daubechies) yield better performance?
- Is there a better way to do the final discrimination between quasi-periodic sources and isolated events?
- Would a time-warping construction on our bases yield better results for irregularly spaced data?

Further applications

- We are currently applying this method to a set of time series from the PanSTARRs project.
- Unfortunately, because this is essentially raw output from the CCD, the preprocessing was not completed in time for this talk.
- We look forward to applying this method to other datasets in astronomy and other fields.